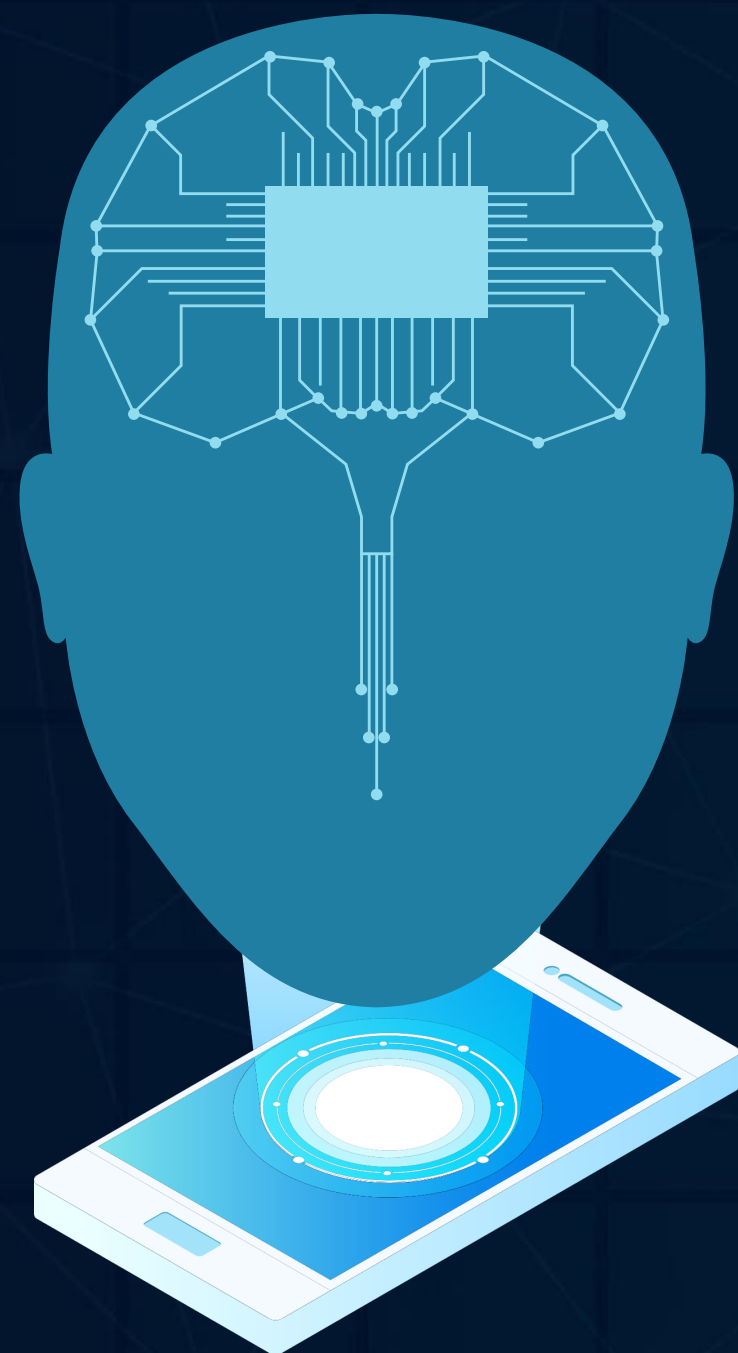
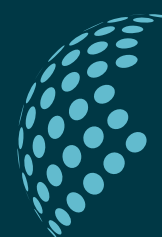


WEB SCRAPING AND AI TRAINING IN THE DIRECTIVE 790/19



CHIARA GALLESE



MARIE SKŁODOWSKA CURIE
POSTDOCTORAL FELLOW
DEPARTMENT OF LAW
UNIVERSITY OF TURIN

Project 101108151 — DataCom — HORIZON-MSCA-2022-PF-01

A new EU Framework for an Ethical Re-use of Health Data

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.



Funded by
the European Union



TABLE OF CONTENT

01 Generative models

02 Web scraping

03 Directive 790/19

04 GM training in the Directive



GENERATIVE MODELS

Generative models use various neural network architectures and training methodologies to generate images, code, or text that closely resembles human-created content.

The peculiarity of these tools is that they employ a large number of data to train their models. In fact, to reproduce a certain style or subject, they need many different examples, usually labeled and filtered by humans. The training dataset size often correlates with model performance, with larger datasets generally leading to better results.

They usually employ web-scraped data in their training corpus.

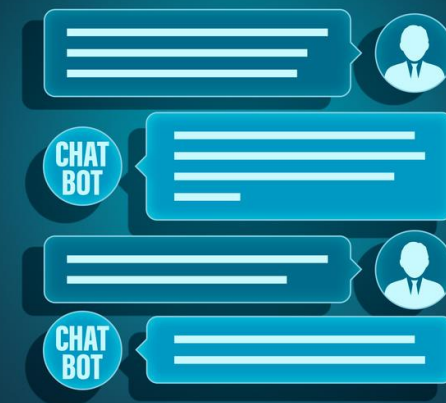




WEB SCRAPING

PUBLIC DOMAIN DATA

Part of the content scraped from the web is of public domain, originally or after the expiring of the copyright



COPYRIGHTED DATA

Scrapers cannot distinguish between copyrighted data and non-copyrighted data when downloading a significant amount of information from the web.

PERSONAL DATA

Data available on the web might be personal data covered by GDPR

ILLICIT CONTENT

Scrapers cannot distinguish illicit content when downloading data, it must be reviewed by humans

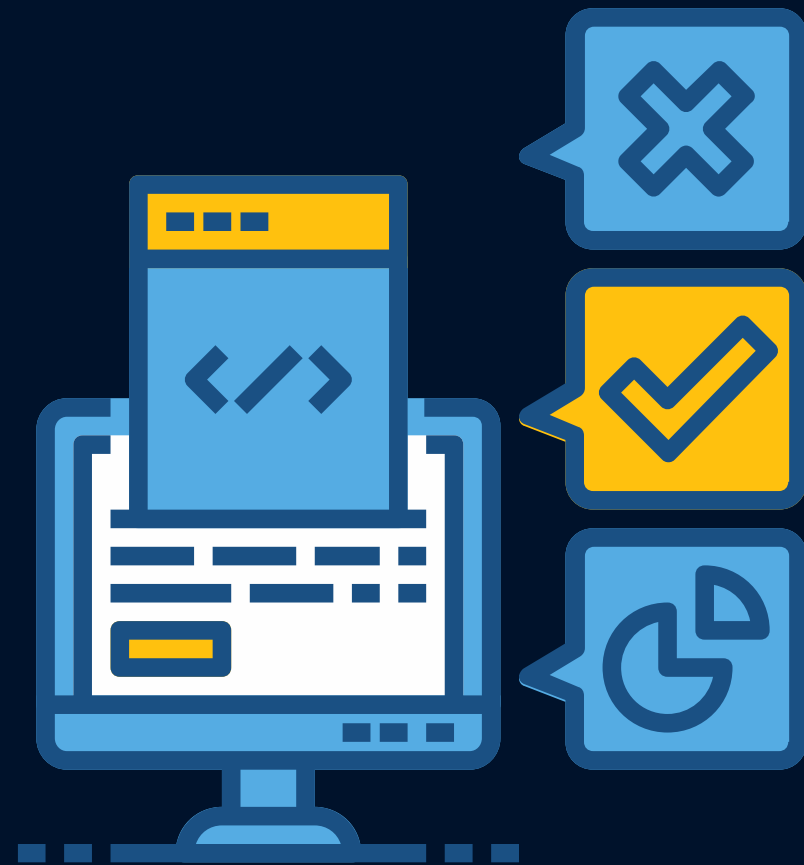


DIRECTIVE 790/19

According to Recital 7 of Directive 790/2019, “New technologies enable the automated computational analysis of information in digital form, such as text, sounds, images or data, generally known as text and data mining. Text and data mining makes the processing of large amounts of information with a view to gaining new knowledge and **discovering new trends possible**”.

Article 2 defines TDM as “any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations”.

TEXT AND DATA MINING



In Computer Science, the practice of **identifying patterns and relationships in vast amounts of data** is known as data mining, referred to also as “knowledge discovery in databases”.

To **computationally find and extract knowledge from unstructured text** is called text mining, also referred to as “knowledge discovery from text”.

ART. 2



Automated computational analysis of digital material performed through analytical techniques

Gaining new knowledge and discovering new trends

Generating new information



WHAT IS NOT TDM



NON-ANALYTICAL TECHNIQUES



OPERATIONS THAT ARE MERELY
TRANSFORMATIVE OF PRE-
EXISTING MATERIALS

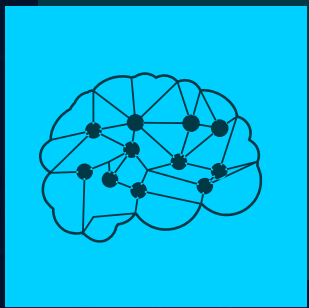


OPERATIONS THAT DO NOT
GENERATE NEW KNOWLEDGE

GM TRAINING IS NOT TDM

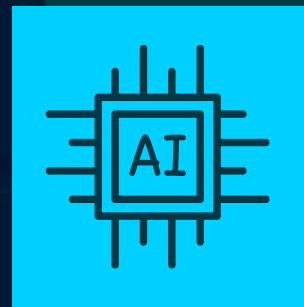


GM TRAINING IN THE DIRECTIVE



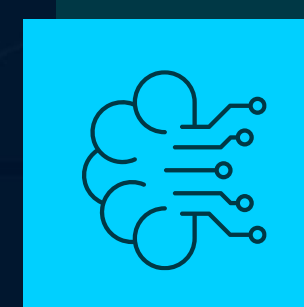
AI COPYRIGHT

GM is excluded from the directive but ordinary intellectual property rules are still in force



WEB SCRAPING

Therefore, there is no web scraping exception for AI training in the current EU legal system



FAIR USE


Although AI training for research purposes might fall under the fair use exception, its use for commercial purposes is not allowed





CONCLUSION

Companies such as Open AI, Midjourney, etc. need to fully comply with the provisions of copyright laws, without the exceptions provided by the Directive, if they want to provide their services in the Member States, and this means asking permission from the copyright holders.



THANKS!

Do you have questions?

chiara.gallese@unito.it

<https://www.datacomproject.eu>

<https://www.aiandlaw.eu>

